Final Project Report

——COVID-19 Research

Weijia Hu

1. Introduction

- Objective

  In this project, we're going to analyze:

  ✧ Until 10/06/2020, which states contribute most to the new cases and deaths?

  ✧ Is there any association between the number of hospitalized patients and the number of deaths for COVID-19?

  ✧ After the National day holiday, did the number of daily new cases increase?

- Background

  COVID-19 is absolutely the biggest issue in 2020 and it affects a lot of people in the world. In the United States, the virus is still spreading.

  In my project, I'm interested in the new cases and deaths of COVID-19. The new cases indicate how fast the virus spread and the deaths indicate how serious this disease is.

- Source of data

  In this project, the data are from The COVID Tracking Project. This project is a volunteer organization launched from The Atlantic and dedicated to

collecting and publishing the data required to understand the COVID-19 outbreak in the United States. The investigators of this project would update the data each day. The date in my data set ranges from 01/22/2020 to 10/06/2020.

2. Method

- Use bar charts to assess number of cases and deaths among all states.
- Use the usmap package to draw maps in order to visually present the data.
- As for the associations between hospitalization and death & between holiday and incidence of disease, line charts were used to examine them.

3. Results

- Summary of the data
  - ✧ There are 12138 rows and 43 columns in the raw data set.
  - ✧ The investigators assigned every state a data-quality grade based on their evaluation of the completeness of states' reporting, I removed rows whose data-quality grades are D, which I think the data are not reliable. I stored these data in the data set named "cov_d". There are 400 observations in "cov_d" data set. Let's check states in this data set in recent days, in which the public health systems of Covid-19 are considered good enough. The states are American Samoa, Northern Mariana Islands and US Virgin Islands and they are not in the mainland USA.
  - ✧ After removing some observations, there are 11738 observations left.

date	state	death	deathIncrease
Min. :2020-01-22	Length:11738	Min. : 0	Min. :-213.00
1st Qu.:2020-04-25	Class :character	1st Qu.: 74	1st Qu.: 0.00
Median :2020-06-20	Mode :character	Median : 500	Median : 4.00
Mean :2020-06-19	NA	Mean : 2075	Mean : 17.23
3rd Qu.:2020-08-14	NA	3rd Qu.: 2054	3rd Qu.: 15.00
Max. :2020-10-06	NA	Max. :25536	Max. : 951.00
NA	NA	NA's :749	NA

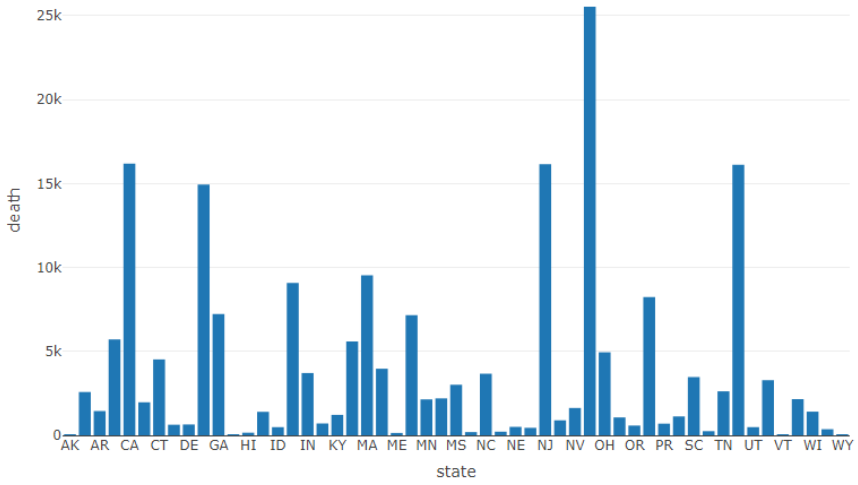
date	state	positive	positiveIncrease
Min. :2020-01-22	Length:11738	Min. : 0	Min. :-7757.0
1st Qu.:2020-04-25	Class :character	1st Qu.: 1936	1st Qu.: 36.0
Median :2020-06-20	Mode :character	Median : 15326	Median : 244.0
Mean :2020-06-19	NA	Mean : 55734	Mean : 633.5
3rd Qu.:2020-08-14	NA	3rd Qu.: 62194	3rd Qu.: 711.0
Max. :2020-10-06	NA	Max. :828461	Max. :17820.0
NA	NA	NA's :99	NA

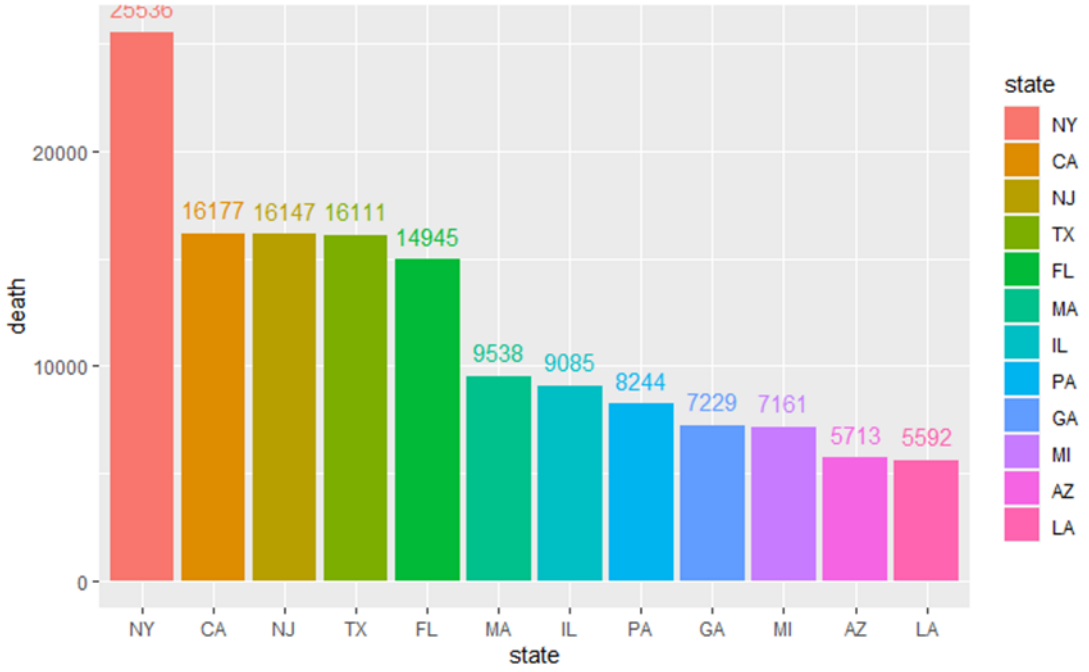
This is the descriptive statistics of the data. The date variable started at 01/22/2020. The first case in USA was found on 01/21/2020. So, our data made sense. Death, deathincrese, positive and positiveincrease ought to be positive numbers. However, the mininmum of deathincrease and positiveincrease are -213 and -7757, respectively. It didn't make sense. We need to remove these incorrect values.

- The bar charts

    1. Total death number by states by 10/06/2020.

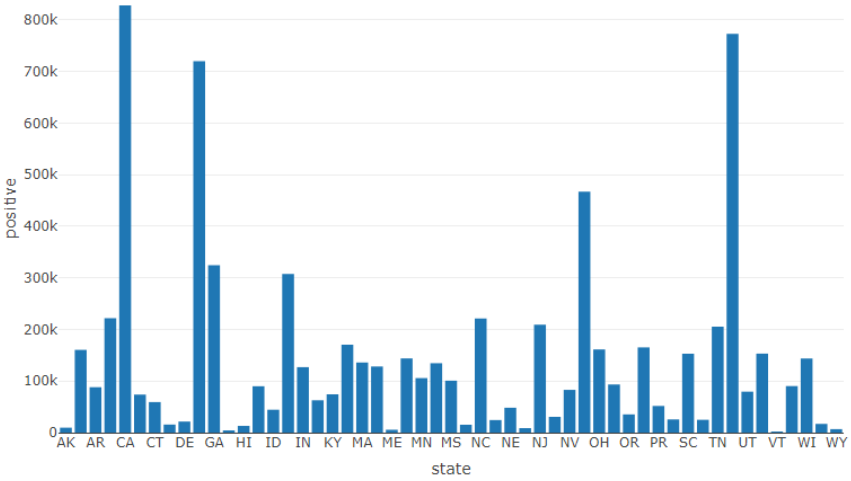
        First, take a look at the first bar chart. There are several states being in really terrible conditions. I selected states whose number of death is greater than 5000. Then sort them by their death descending. Among them, the New York has the highest total deaths number in the USA. The state of California, New Jersey and Texas have the similar numbers and are ranked second, third and fourth respectively.

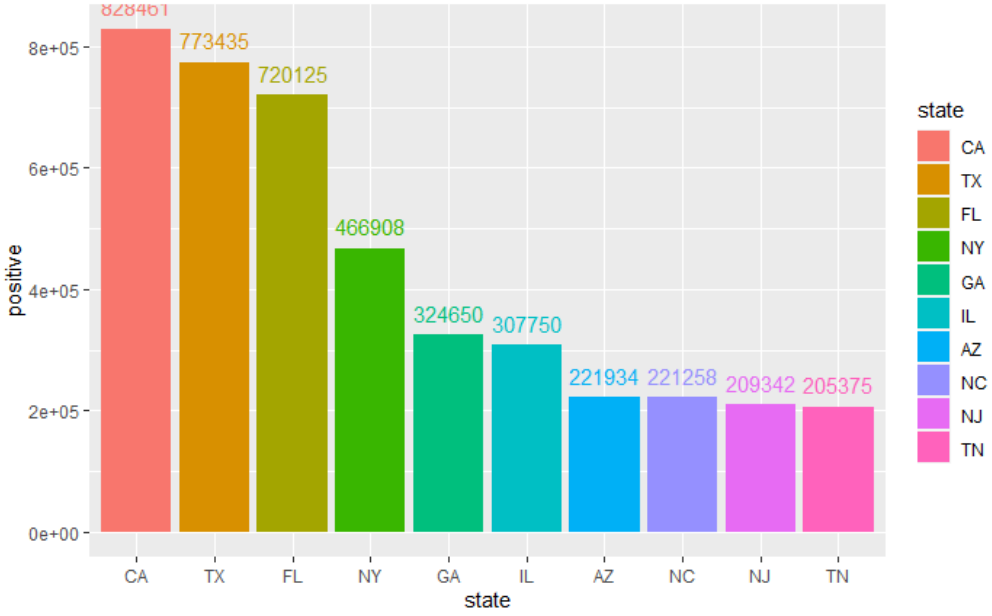




2. Total case number by states by 10/06/2020.

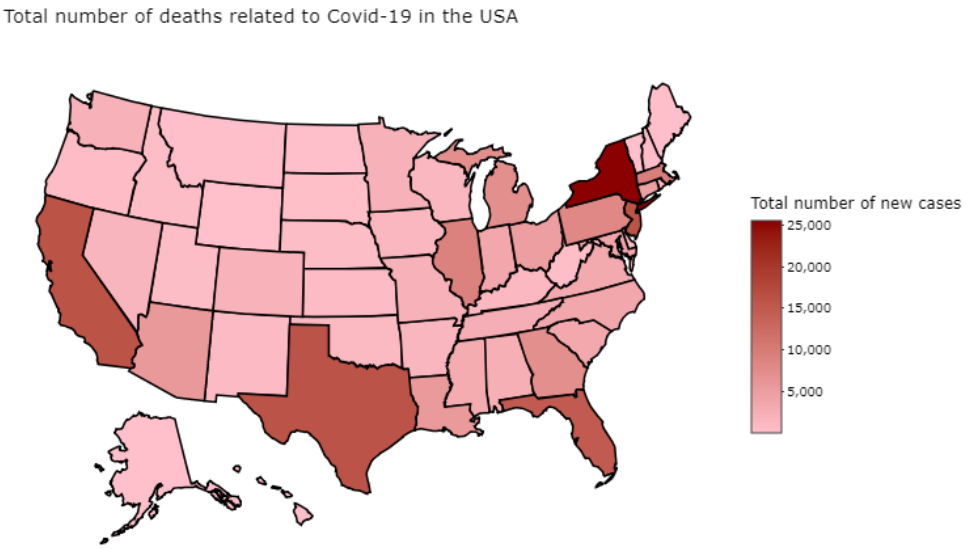
   Take a look at the first bar chart. There are several states being in really terrible conditions. I picked states whose number of cases is greater than 20,000. Then sort them by their number of cases descendingly. Among them, the California state has the highest Covid-19-related cases, which is what we all know. Interestingly, the state of New York is only ranked forth, while it has the most deaths.



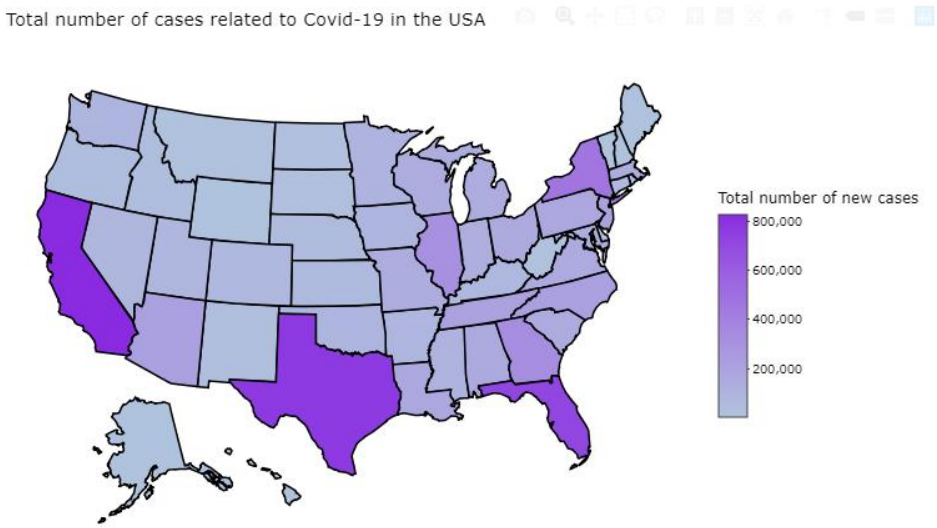


- The maps

1. Total death number by states by 10/06/2020.

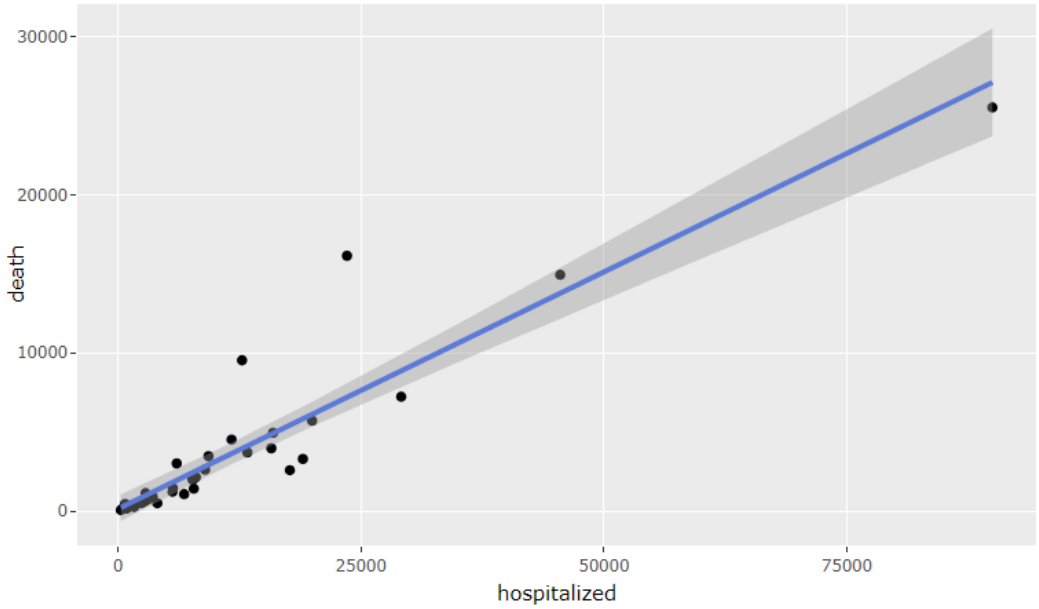


2. Total death number by states by 10/06/2020.



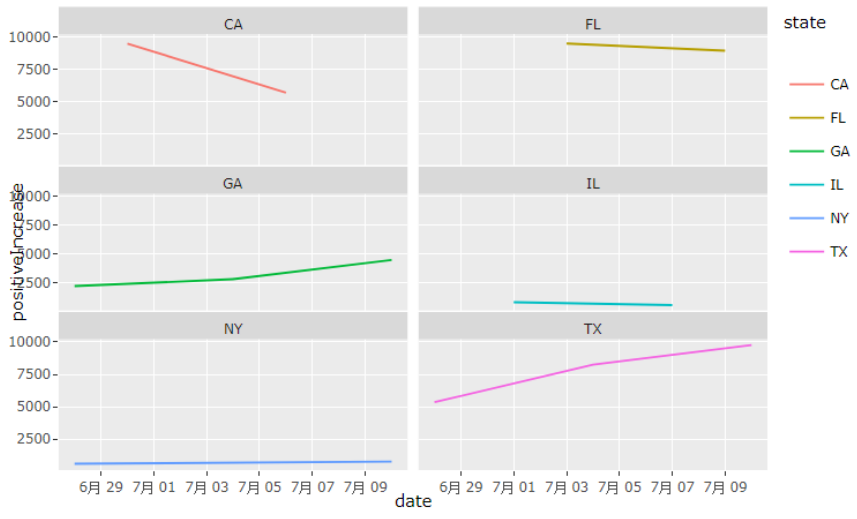
- Hospitalization and death

  In my dataset, many states didn't report their hospitalization data. Or some states started to report them very late. It will affect the result. Therefore, the result is not accurate enough. Also, in this dataset, there were no data about the survival rate. If we could access the survival data, we could improve our model.



- Holiday and new cases

  As we all know, this summer, there was a peak of new cases. The National day holiday was in the middle of this summer. Did the National day holiday affect the increase of new cases of COVID-19? In order to explore it, I would use data of a week before and after National day to display the trend. I selected top six states in highest new cases: California, Texas, Florida, New York, Georgia and Illinois.



- Conclusions
  - ✧ The maximum of the daily increase cases is 17820 in a certain state. The mean of the increase for every single state is 879 cases per day. The maximum of the daily increase deaths is 951 in a certain state. The mean of the increase for every single state is 24.83 deaths per day. Based on this, the Covid-19 disease could spread fast but not cause so many deaths (but we still need to wear masks and keep social distance to prevent it!).
  - ✧ According to the scatter plot and its smooth line, the number of deaths and the number of hospitalization have a positive relation. Notice that the confidence intervals are getting larger. It's because only some states have very high number of deaths.
  - ✧ I selected a week before and after the National Day to analyze this question. Only in Georgia, after the National Day holiday, the daily increase cases become more. So, I don't think there is any strong relationship between them.